Review of 'Towards variance-conserving reconstructions of climate indices with Gaussian Process Regression in an embedding space'

The manuscript by Klockmann et al. introduces a new technique for climate index reconstructions from proxy data: Gaussian Process Regression (GPR) in a modified input space named embedding space. The new method is compared with classical principle component regression in pseudo-proxy experiments of varying complexity (PPEs). The PPEs suggest that the new method could be superior to established approaches in reconstructing indices with substantial variability on multi-decadal and longer timescales such as the Atlantic Multi-decadal Variability.

The new method and the results are definitely interesting and worthy of publication, albeit I do not exactly see how the manuscript fits into the list of GMD manuscript types. It does not fit into my understanding of the scope for 'model description papers' because while it describes a new method, the implemented model still seems in a rather experimental stage and the README of the code in the Supplement explicitly states "The scripts are taylored to use the provided test data, i.e. they are not written in a general form that would allow to use them with any kind of suitable dataset, yet." This will make it very hard for readers to use the new method outside of the presented PPEs. If it is designed as a model description paper, at least the model must be given an explicit name and version number following the GMD guidelines and some more effort should be put into making the code usable for others. Since the paper develops a new method and is not related to model improvement it is also not directly a 'development and technical paper'. Therefore, I ask the authors to clarify how the manuscript fits into the GMD manuscript types and adapt it accordingly.

In addition to these general considerations, I have a few major issues I kindly ask the authors to address before publication and additional specific comments listed below.

Major issues:

1. Introduction: The paragraphs l. 27-79 describe the AMV-related research fairly extensively and in my opinion much longer than necessary for a model description/development paper. In contrast, the final part (l. 80-89), where the new method is introduced, is a bit short to help me understand the authors thought process in selecting and developing the described methods.
2. Sect. 2.2: The model description varies between long descriptions of general GPR theory / modeling options and fairly short parts on the selected solutions, the motivation behind these choices, and implementation details. I would prefer to focus more on the specific choices for reconstructing the AMVI and why these choices are made. For more specific questions arising from the method description see below.

3. How is σ_n handled for the AMVI? Is the AMVI just given by f(z) or is ε also added in observed and reconstructed AMVI?

4. Sect. 3.2 / 3.3: In the evaluation of the PPEs with noisy pseudo-proxies, the authors focus on the ensemble mean time series across all randomized experiments. I do not see why this is a useful quantity for evaluation. It is not a quantity occurring in reality since the authors correctly state that in real-world applications we only have one realization of pseudo-proxies. Therefore, the ensemble members should be evaluated separately (as each of them is a single realization which could occur in reality) and then the mean and spread of the evaluation measures should be reported and analyzed.

5. As the authors state, GPR is a Bayesian method. Thus, it naturally produces uncertainty estimates through sampling of the posterior distribution. Currently, only the posterior mean is evaluated throughout the manuscript if I see it correctly. Uncertainty quantification is an important part of climate index reconstructions and has been the subject of intense debates over the last decades. Therefore, I would like to see some evaluation on how useful the uncertainty estimates provided by the posterior distribution are.

6. Sect. 3/4: While the authors report several situations where one or several of the reconstruction methods give unreasonable results or fail to reconstruct the underlying truth, explanations for why the models are better in some aspects and worse in others are mostly missing. The authors speculate on some potential reasons but I would like to see some more sensitivity tests to give the reader a better feel for the strengths and weaknesses of the different methods.

Specific comments:

- l. 12: Please reformulate one of the 'relevants'
- l. 17: The last 1000-2000 years are normally named the 'Common Era' and not the 'preindustrial period'. 'Preindustrial' could lead to confusion with simulations using fixed preindustrial boundary conditions. Focusing here on the Common Era as a 'must' seems a bit arbitrary since also other periods would be of interest.
- l. 58: Which non-linear methods have shown promise? Is non-linearity really the main advantage here or could other factors also be important?
- l. 103-105: Which climate variable do you use to construct the pseudo-proxies (e.g. SST, surface temperature, near-surface air temperature)?
- l. 109-111: How do the SNR and the construction of pseudo-proxies compare to other pseudo-proxy studies?
- Figure 1a: From the color scale in (a), it is very difficult to distinguish the correlation of the different records. Maybe you can improve the color scale.
- Figure 1c-e: Over which period are the cross-correlation, STD radio, and embedding distance computed?
- Figure 1d: The 10^0 could be removed.
- l. 132: Are Matern functions really 'very complex' kernels?

- l. 139-140: The inference strategy is described very briefly here. Some more explanation could be useful for readers not that familiar with Bayesian inference
- l. 160: Is there a reason why the abbreviation SVGP is not adopted in the manuscript and 'sparse GP' is used instead?
- l. 163: What is the 'Adam Optimiser'?
- l. 171: How is the AMVI formulated in proxy space since it is defined over a different spatial scale than the individual proxies and how does GPR-based climate index reconstruction work when the GP is formulated as function of the proxy values (=temperature?)?
- l. 177-192: I struggled to understand why the embedding space needs to have the given dimension and the idea of how the embedding space is constructed did not become clear to me until I finished reading Sect. 2.2.3.
- l. 203-205: What are the properties that need to be fulfilled by the distance matrix / distance measure? Why is a positive correlation between records needed between the records?
- l. 209: I guess that equidistant coordinates perform worse because than all records influence the AMVI roughly equally whereas for CC-based coordinates records with a high correlation with the AMVI become more important. But why does including the SR improve the result compared to just using CC?
- Equation (3): Is RA = SR?
- l. 213-218: The description was a bit short here for me to really understand what is happening and why.
- Equations (4/5): The Gaussian kernel functions lead to very smooth (infinitely differentiable) functions, likely much smoother than most processes actually observed in climatology. Does this lead to overly smooth predictions on certain timescales and did you test the procedure with kernels that lead to less smooth posteriors? It also might be useful to write down the final covariance model as an equation.
- l. 230: Is σ_n^2 the same as a nugget effect in statistical modeling? If so, it might make sense to mention it here for the statistically-inclined readers.
- l. 235/236: Do you have an explanation for why this slightly unusual formulation (two kernels acting in time but only one kernel acts in the "embedding space/distance") performs better than models without $k1$ or with separated kernels acting in time and embedding distance? Does this indicate that the system (AMVI) is better described by two characteristic timescales instead of one similar to two-box energy balance models outperforming one-box energy balance models in predicting sea surface temperatures?
- Equations (6) - (9) did not help my understanding. Either they should be embedded better in the text / explained better or they could be removed.
- l. 259: There is a typo in 'obtain'

- l. 299/300 (and similar parts in subsequent sections): The difference between the MPI-ESM- and CCSM4-based results could be explored a bit further. What are the main differences between the simulations that might explain the differing behavior of the reconstruction methods?
- l. 313/314: Is there an explanation for why the sparse GP performs better for noisy than for perfect pseudo-proxies on multi-decadal timescales?
- Fig. 4: This is an interesting figure to explain some of the differences between methods displayed in Fig. 2/3 but unless I missed something, it is barely discussed in the text.
- l. 359: Could the overestimation of variability in periods with few available proxy records be explained by relying too strongly on a small number of proxy records which tend to be more variable than the AMVI due to integrating over a smaller spatial scale? This could maybe be tested by comparing hyperparameters fitted separately for periods with high and low record availability.
- l. 393-398: Is the introduction of a 'noise variance' parameter similar to error-in-variables approaches for frequentist regression models?
- l. 408: Is the model really using one tenth of the available training data if you use every tenth time step but also an (optimized) subset of the original locations?
- l. 416-418: Can you expand on these length scales and magnitudes? What are expected values and where do the parameters rank in the range of reasonable values?
- l. 444: The GPR-model seems to handle white noise proxies very well, in parts due to the inclusion of the parameter σ_n. What would happen if the proxy noise would be auto-correlated? Is there a way to adapt the model accordingly?
- Conclusions: Since this paper develops and tests a new methods for climate index reconstructions, it would be very useful for the reader to get some more guidelines for future applications of the method and how it might be applied to other indices.